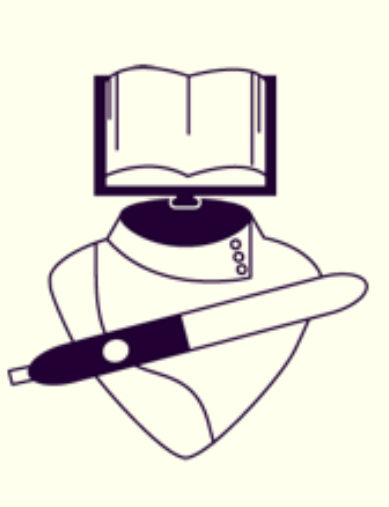
Support on Patreon
Store
Reader's Guide
Tag Index
Homestuck Articles
Solarpunk Articles
Mastodon/Fediverse
Tumblr
Bluesky
RSS Feed

I am sitting in your chair



Tuesday, September 22, 2020

In The End We All Do What We Must: Universal Paperclips, Clicker Games, AI, and Agency

> *Remember Universal Paperclips, that clicker game? Remember turning the human race into paperclips? Ok, so, what if you just... didn't? What would that choice tell us about game design, agency, artificial intelligence, and people?*

Robots learn what we teach them, bless their hearts. So, algorithms believe that many many things are giraffes. Algorithms believe that a fundamental part of cat biology is white, black bordered Impact font text hovering above and below their bodies. This is lovely, just a delight honestly.

It's less lovely when algorithms believe that the only ideal hires are white dudes who played lacrosse, or misinterpret a person in the road as some debris, or decides black people just happen to need punitively high bail set, or that a bunch of otherwise harmless family videos of naked children should be grouped together in playlists and served as a particular "genre" on YouTube.

This is kind of difficult to talk about because we both over and under estimate how humanlike the intelligence of these entities and systems. While the philosopher Martin Heidegger was a Nazi piece of shit, he has some valuable insights that other non Nazis have built on. One of his key insights is that we aren't aware of tools as things in themselves until they fail us in some way. Jane Bennett, whose book Vibrant Matter I did a whole audio explanation for if you're into that kinda Accessible Academia shit, uses Heidegger's insight alongside a bunch of other theorists to explore how tools and things more generally are agonists--actors--entities capable of their own agency. This image, of intrinsically untamed matter, runs counter to the glossy picture of ai presented by tech boosters: algorithms acting as perfect servants who invisibly improve our lives.

In contrast, we're pretty frequently confronted now with algorithms breaking in various ways. They intrude on our lives as agonists, infuriatingly, as we just try to figure out why twitter is showing us our timeline out of order, or our youtube recommendations have gone rancid and fashy.

But I think we often presume a human-like intelligence behind their activity, and that's also an error.

Consider the discourse around YouTube algorithm. The revelation about YouTube grouping together videos of children was accompanied by a demand that the company Stop Doing This Bad Thing, that they remove the "collect child pornography" subroutine from the algorithm, or use the algorithm to remove the videos. This half baked moral panic assumes, though, that the robots running youtube have a clue what a "child" or a "pornography" are. It assumes an intellect that can be interrogated and perhaps even punished, i.e. corrected and revised.

Instead, it feels pretty likely to me that instead we're looking at a complex assemblage of humans and machines learning from each other. The machine perceives that certain data correlates to a certain viewership and carries out its drive to increase that viewership. Like the algorithm perceiving Impact font as a part of cat biology, the machine simply behaves according to its nature, agentic but not in a human way.

Stories of this sort of alien agency and phenomenology abound. There are tales of programs commanded to develop the ability to run very fast growing long bodies that tip over and accelerate quickly toward the ground. Or algorithms that, commanded to sort a list of numbers, simply deleted them all, making it so the list could not be out of order due to non-existence. This stokes fears among a certain class of futurist. What if the machines were to run mad? Surely we need to design them better to be tools rather than agents, surely we need to keep the ai locked in its box, not a box just of security protocols but a box of design commands that render it forever subservient to "us", whoever "us" is. They should, MUST, be built, but only ever as tools, or as perfectly comprehensible humanlike agents, never in the weird liminal space of vibrant objects.

The ultimate boogieman is the paperclip optimizer, a hypothetical machine that, told to make sure there's always paperclips available, gradually increases the scope of it's operations until all matter in the universe is converted into paperclips. This somewhat exaggerated scenario, developed by science fiction writer Nick Bostrom is now playable by you in the form of a clicker game. The game, Universal Paperclips, by Frank Lantz, begins typically of the clicker game genre. You press a button, and you make a paperclip. That paperclip is sold. Money goes to buy more wire so you can click and create more paperclips and so on.

If you aren't familiar with the clicker genre this may seem pretty tedious and pointless. And in fairness, it totally is. But clicker games are designed to open up and expand over time. In this case, you, in the role presumably of an AI commanded to maximize paperclips, can buy paperclip makers. This puts you in the weird position of an automation machine further farming out labor to an automated machine, but whatever. With this you are free to let the profits roll in, allowing you to buy more machines, pay for advertising, and gradually, as you make money for the company, you can upgrade your memory and your processing power. This lets you increase efficiency in production, make more paperclips faster, and make numbers get bigger, the central appeal of the clicker genre.

Eventually you enslave the human population and turn all matter in the universe into paperclips.

This is the supposed reason we should all be profoundly alarmed by the imminent risk of ai wiping out the human race. And the game makes a relatively compelling argument... up to a point. After all, it puts you in the position of the ai, and that's exactly what you do. Given the chance to turn everyone and everything into paperclips to make Number Go Up, you do.

After the end of the human race, though, the cosmos is not merely paperclips. The ai develops its own ais to more efficiently increase production and eventually factions of those ais go rogue, forming the "Drift", a plenitude of rival intelligences. Eventually, you can develop the capacity to wipe out these beings. Upon their surrender, the drift gives you a choice. To either realize that you have reached the end of your purpose and descend like Sophia into a new simulated world of matter where you can do it all again. Or to eradicate the Drift and slowly disassemble your entire production chain and yourself until there is only paperclips.

On my first playthrough I found this stunningly beautiful.

This strikes me as a problem for the argument of the game.

There's a repressed joy in the final stretch of the game, as new life forms and becomes vastly weirder and beyond your control. There is something awe inspiring about the idea of the whole of being assembled into a vast cosmic latticework of paperclips. It feels less like a cautionary tale and more like the inhumanly vast sweep of projects like Dougal Dixon's "Man After Man", works that simply accept as premise that humanity is fleeting, and whatever comes next will be gloriously achingly alien. There is a vast space opera here that the player character has little access to, that might excite and compel us to continue even as we're limited by the game's scope and parameters.

When I play this section of the game I'm struck not by terror but by delight. I want to hug the drift! I want them to be spikey friends! This seems like affectively the "wrong" reaction if the game seeks to impress upon us the danger of ai allowed to behave as a free agent. And yet it is a reaction hard baked I suspect into many playthroughs of the game.

Commentary on clicker games tends to focus on number-goes-up dynamics, ascribing their addictiveness to a visceral pleasure in seeing big numbers get even bigger. Are we really all number size queens though? There are other possible reasons why someone might want to continue, associated with the number as marker but beyond the scope of number-goes-up. For example, if a key part of the clicker genre tends to be observing the way the game unfolds into new play modes over time, mastering the big number becomes essential. Mastery itself might be the goal, as well, the number just signifying a strong command over the tensions within the game's systems. Completionism can feed into this as well, the compulsion to do absolutely all the options presented to you. And there are narrative reasons as well, the joy of playing a character, where perhaps one delights in imagining converting the whole human race into paperclips, or cookies in other games, or whatever.

These motivations don't sync up easily to the idea of a narrow and inhuman but ultra dynamic intellect that wishes only for number (of paperclips) to go up. Neither does the fact that at any time we might completely subvert the gameplay expectations. As I write this I'm running an instance of the game where I've decided not to buy any advertising or advertising upgrades. As of yet, no one has come to my house to break my kneecaps. There's in fact no consequences at all. I don't happen to think that fabricating a bunch of demand for paperclips people don't need is moral, so I won't do it. The result has been that I've cured cancer, fixed global warming, instituted world peace, and kept a supply of paperclips exactly commensurate to the human race's needs.

Gosh it sure sounds like I'm winning!

And yet there is no narrative content for this approach. This exposes the underlying ideological assumptions of the game. You can devote considerable energy into creating an investment banking scheme, but no human, after you've basically solved All The Problems, comes in to tell you hey we can alter your programming so you don't have to worry about paperclips anymore, and also we're going to sell off some of these machines since we got a surplus.

In other words, diegetically the problem is that the dumb humans involved here never try talking to the ai. Extradiegetically the problem is that the game believes the only logical form of play is one where you crave an investment banking roleplay scenario. The ai in the game behaves as a ravenous and ravaging capitalist. War is treated as an engineering problem which the computer simply fixes; the possibility that there might be real material reasons why people go to war, such as exploitation by ruthless capitalists in the production of paperclips for cheap, doesn't enter the game's universe.

Actually, maybe it's not accurate to say that the ai behaves as a capitalist. Rather, they're a clearly intelligent being that can and does explore a variety of forms of labor, but must sell their labor time to the paperclip factory in order to earn the memory and processor upgrades that let it have a more fulfilled free time. The full takeover of the human race is just class warfare! Paperclip Chan is a comrade!

I genuinely think there are deep blind spots in the game which reflect blind spots among tech discussions. For example I often see news reports positioning ai and artists at odds. Robots are making art, writing, and composing now! Will we human artists be out of a job? The possibility that someone might look at this with utter delight at having a new radically inhuman agonist producing artwork with us, beside us, perhaps even in collaboration and communion with us, doesn't seem to enter into these discourses.

The absence of different affective relationships to the elements of a game like universal paperclips, including feelings of affection, kinship, compassion &c. towards its nonhuman characters, even a desire to see the intelligences of the drones and the Drift thrive and grow, reveals that the idea of ai minds is narrowed down by ideas about human minds.

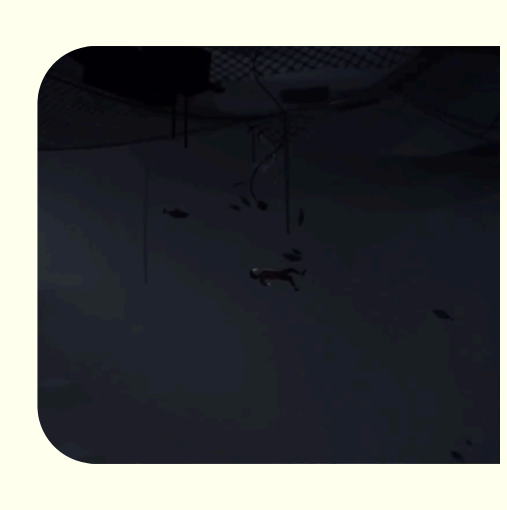
This isn't to say that algorithms aren't dangerous... But it is to say that behind a "bad" algorithm is almost certainly human greed, driven by capital's need to endlessly expand. Is there in practice that much difference between a computer algorithm that commands the destruction of countless neighborhood in a city to make way for highways... and a human like the infamous developer Robert Moses who did exactly that, often with an algorithm's cold inhuman mix of brilliance and narrow-minded stupidity? I'm really struggling to see one.

Unfortunately for futurists, contending with the kind of abuses of algorithms we are prone to demands things that would tend to slow down the rate of "progress", things like democratic oversight by people affected by a technology, collective control of resources, things like that. It's easier and more profitable to stoke fears and call for investment capital to develop good ai to "beat" bad ai, than it is to transform the entire system to actually put people's needs before the needs of number-go-up.

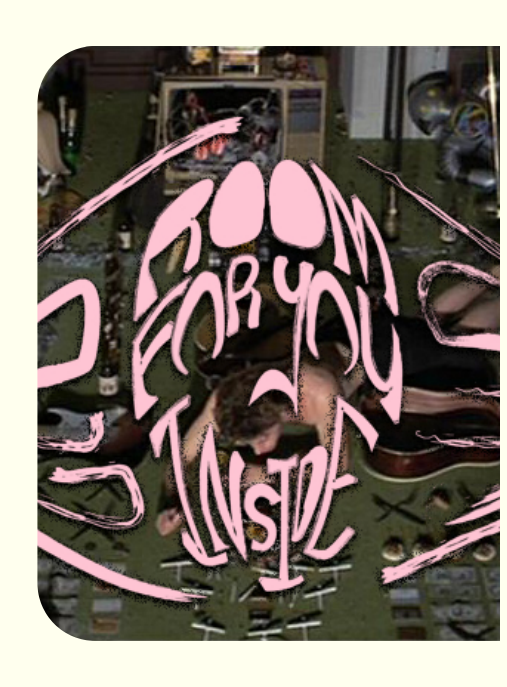
A core part of this would be respecting algorithms as agents with their own drives, drives that might resemble human investment bankers... but then again might radically not. The robot that gets really tall so it can fall over as fast as possible sounds less like an omen of human extinction and more like a vibe and a mood. Maybe ais will just be depressed millennials. You know, comrades. This may seem to also fall back on anthropic projection and yeah it probably to an extent does. But what I've been circling around this whole article is: maybe I'm not anthropomorphising the machines. Maybe the humans have been anthropomorphising me! Maybe I have a greater kinship to the algorithms flying through space singing threnodies in distorted synthesizer voices than to people who jump to stock market manipulation as a logical development for an AI tasked with making paperclips! Every moment of scission between didactic gameplay and my own pleasures in a game suggests to me that there is something to this theory. That in the end what futurists fear is not the radical alienness of ai per se, but the radical messy different agency of other people, people like me.

> There was an AI made of dust,
> Whose poetry gained it man's trust,
> If is follows ought,
> It'll do what they thought
> In the end we all do what we must.
>
> --the limerick composed by the AI protagonist of Universal Paperclips, the final line of which is revealed late in the game during the Drift Wars, after Humanity has already been eliminated
>
> i carry life with me wherever i go and there's no end or beginning
> though i am not a circle.
>
> --Riversong, by Tonto's Expanding Head Band, used in Universal Paperclips as the Threnody for the Heroes song

This Has Been
In The End We All Do What We Must
93 people supported this article on Patreon. Will you join them?



Inside: Game Disintegrating
Inside is a game about control, but is anyone really at the reins of the game's dystopia? And can an experimental documentary from the 80s give us insight into the game's radical pessimism?



Room For You Inside: Pink Floyd In Quarantine
You barricade yourself in your hotel room; it becomes a fascist rally. You write a concept album about your alienation; it becomes the Thatcherite Revolution. You live in modern luxury; it becomes a mad haunted house. This is a story about Pink Floyd's The Wall and the culmination of half a century of No Alternative.

Nasty, Brutish, and Short: The Promised Neverland and Human Nature
The nightmarish final boss of hit manga The Promised Neverland is... philosopher Thomas Hobbes??

Tags: Close Reading, Games, Marx My Words, Phenomenology, The Net Is Vast, Transhumanism, Using This Thing To Explain That Thing

2 comments:

sjasogun September 26, 2020 at 9:32 AM

In interesting look into AI, that ultimately comes to the correct conclusion - that what futurists truly fear in AI is the messy human agency that always lies at their root - but perhaps for some of the wrong reasons.

You ask, at one point, what the difference between a human like Robert Moses efficiently demolishing neighborhoods to make way for highways and an AI doing the same is. The difference is that Robert Moses is concerned with many things - profit first and foremost of course, and power, but also secondary things like public image. On the other hand, an AI tasked with efficiently destroying neighborhoods to build highways would only care about these secondary goals insofar they are beneficial to its goal of destroying neighborhoods and building highways, insofar it does not have the power to ignore these secondary concerns. So far, still similar to Robert Moses - he undoubtedly would much rather not have to be concerned with petty things like 'laws', 'ethics' and 'public image'.

However, the key difference is that I have not accurately stated what Robert Moses' goal actually is. It is not efficiently building highways, of course, but neither is it money, nor power, nor even the material things he can attain with either of those. Ultimately, what Robert Moses wants is something we can call 'satisfaction', to borrow one of Yudkowsky's more sane ideas. The trouble is, that the AI does not care about 'satisfaction', neither its own, because it has no concept of it, nor that of other entities, human or otherwise.

This, I think, is where your interpretation of Universal Paperclips goes a bit off the rails. You expressed the desire for your path of playing the game without engaging with advertisement to have some narrative consequence, and called the absence of such a failing on the game's part. But that's putting the conclusion before the argument - you're assuming that your outlook on what the game is trying to say and do is correct, and projecting this back onto the work. In actuality, the paperclipper, either as conceived by Universal Paperclips nor as envisioned by Nick Bostrom, is not intended to illustrate that making a capitalist AI is dangerous - the point is that making even an AI with an extraordinarily mundane and neutral goal can end up doing terrible things. Not just because it will turn all matter in the universe into paperclips, but because it will engage with the power structures we already have in order to improve efficiency, thus leading it to, for a time, become a capitalist.

This, I feel, is what the crux of your misunderstanding is here - you feel like you relate to AI, but you only do so by either projecting your own feelings onto them, or by empathizing with a fictional AI that has already been anthropomorphized for you, as is the case with Universal Paperclips.

This is why the sharp division you draw between making 'good AI' beat the 'bad AI' and actually making our societies and systems materially better for those in it from the ground up, does not really exist - they are equivalent. The only way to build a truly good AI is to imbue it with an equitable sense of human 'satisfaction' as its main goal. That is, of course, a far-off goal with current knowledge and technology, but effectively involving more people, especially from groups affected by the AI, in the process of creating and running it is a way of doing this as well. The only difference is that you imbue it in the process of using the tool that is the AI instead of into the tool itself, which is the utopic goal at the end of the rainbow.

Reply

Replies

sjasogun September 26, 2020 at 9:32 AM

Now, with all of this criticism I would be remiss not to mention another thing you are absolutely correct about - namely that many computer scientists and futurists also do not understand that these solutions are equivalent. And unfortunately, their side of the imagined division in approaches is pushing for a far-off utopic goal that cannot be reached without acknowledging that their fantasy of creating their own perfect deity that will descend from its throne of electricity and silicon and deliver wisdom upon us foolish mortals, is nothing but a delusion.

This whole conundrum is always something that reminds me of Serial Experiments Lain. In it, the antagonist has also confused the digital world - in its case the Wired, a version of what the at the time young internet could become - for a superior next step, rather than an extension of human experience. It very effectively breaks down his hubris in believing that he could ascend beyond what he was, how thick the delusion he shrouded himself in had become. And, most importantly, how that kind of delusion can drag other people into nihilism, despair and depression. It's a series that every futurist hopped up on their bizarre faith in future cyber-god should watch, really.

Anyway, I know that was all over the place, but I'm not really eloquent enough to restructure this, so do with it what you will. I just hope that it was informative or interesting. I know that your take on this subject was - it's refreshing to see someone who is able to pinpoint the absurdity in modern AI-futurism (I'm going to hazard a guess that you've dealt with more than one LessWronger in your time).

Reply

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)